Statistikproduktionsstödet

Mätteknik 2.0

Sammanfattning

Förutsättningarna för statistikproduktion har förändrats över tid och tar sig bland annat uttryck i stigande kostnader och bortfall för direktinsamling, men även i ökad tillgång till andra typer av datakällor med potential för statistikproduktion. Denna förändring är en utgångspunkt i SCB strategi som styr mot en förflyttning från direktinsamling mot andra typer av datakällor. Medan traditionell mätteknik alltid varit starkt förknippat med frågeformulär och metoder som används vid direktinsamling, beskriver Mätteknik 2.0 hur mättekniken kan bidra till utvärderingen av nya datakällor, som ofta inte baseras på frågeformulär men ändå kan innefatta mätproblem.

I princip innehåller mättekniken 2.0 samma centrala utgångspunkter som den tidigare mättekniken:

1) Beskriva datagenereringsprocessen i syfte att ge en förståelse för förutsättningarna kring hur data bildas och eventuella problem. Ett enkelt exempel på en källa innehåller dessa processteg: *en handläggare på en myndighet har en personkontakt med en klient vilket genererar vissa uppgifter, handläggaren interagerar med ett digitalt verktyg för att föra in uppgifterna i myndighetens databaser. Uppgifterna bearbetas internt på myndigheten för att slutligen resultera i datakällan av intresse för SCB.* Även i detta enkla exempel finns det många faktorer som SCB behöver en djupare förståelse kring för att kunna ha tilltro till och tillämpa källan. Mätteknik 2.0 beskriver metoder för att ge den kunskapen genom t.ex. intervjuer med sakkunniga och genomgång av dokumentation.
2) Den andra delen i mätteknik 2.0 handlar om att undersöka hur en källas innehåll passar en på SCB tilltänkt tillämpning. Det handlar om att utreda graden av överenstämmelse mellan källans och statistikens definitioner angående t.ex. population, referensperioden och centrala begrepp.

Mätteknik 2.0 är en del i utvärderingen av nya datakällor, men metoderna kan-vid behov även tillämpas på etablerade källor. Det är dock ingen isolerad del utan bör utföras i samarbete med främst statistisk metod och ämneskompetens.

Introduktion

Detta dokument innehåller beskrivningar av metoderna i mätteknik 2.0: mätteknisk granskning av nya datakällor där data bildats på annat sätt än via frågeformulär (det sistnämnda betraktas som Mätteknik 1.0). Dokumentets målgrupper är mättekniker och andra som är berörda av utvärderingen av nya datakällor, t.ex. i rollen som användare av en ny källa eller som team-deltagare i godkännandeprocessen.

Nya datakällor kan ha olika egenskaper, t.ex. variera från strukturerad, administrativ data baserat på input från individer till mindre strukturerad maskingenererad data av big-data proportioner. Det mättekniska arbetet är mest betydelsefull för den förstnämnda typen av datakällor men kan även bidra till utvärderingen av den sistnämnda.

Nya datakällor kan oavsett typ variera avsevärt vad gäller t.ex. tidigare kunskap, egenskaper, metodproblem, tillgänglig information, aktuell tillämpning för SCB etc. Med tiden kan innehållet i den mättekniska granskningen avgränsas och beskrivas i ännu mer detalj i en ny version av detta dokument.

Utgångspunkten är att mätteknik 2.0 innehåller följande fyra delar:

1) Beskrivning av datakällans syfte, innehåll och sammanhang. Detta är ett brett avsnitt som inkluderar flera typer av innehåll.

2) Datagenerering i källan. Denna del beskriver datagenereringen i källan. Beskrivningen ska inkludera ett flödesdiagram. Här används ”datagenerering” i bred betydelse och inkluderar även den eventuella bearbetning (t.ex. kodning) som sker internt i källan.

3) Tillämpning av källan för ett statistiskt ändamål. Denna del handlar om att beskriva och bedöma hur källan passar ett specifikt ändamål. Det kan handla om hur källans variabler passar SCB:s standarder men även hur källan passar behoven för en viss, utpekad, statistikprodukt.

4) Mättekniska slutsatser. Denna del innehåller bedömningar vad gäller t.ex. mätfel, saknade värden och bearbetningsfel i källa.

Avgränsningar och andra underlag till utvärderingen.

Datakällor kan kategoriseras på olika sätt. En vanlig variant är UNECE:s uppdelning. Den delar in datakällor i tre kategorier utifrån hur data genereras: människa (t.ex. sociala medier), process (t.ex. företags produktionsprocess eller myndigheters administrativa process) och maskin (t.ex. sensorer och signaler). Mätteknik kan generellt bidra med mest till den mittersta typen men kan vara värdefullt även för andra typer. För övriga typer (maskin, människa) finns kvalitetskriterier (Jansson & Japec ”Kvalitetskriterier för statistik baserad på digitala data, 2023). Några punkter från vägledningen till kvalitetskriterierna har lyfts in i metodinnehållet nedan. I vägledningen för kvalitetskriterierna förekommer frågeställningar (betecknas med bokstaven A) och indikatorer (betecknas med bokstaven B). En tumregel är att mätteknik 2.0 kan bidra med värdefull information för att kunna besvara frågeställningarna i A. Indikatorerna i B bör dock metodstatistiker eller dataanalytiker ha huvudansvar för.

Andra utgångspunkter för utvärderingen av en ny datakälla är ”Kvalitet för den officiella statistiken – en handbok” och mallen för DOKBOR.

Del 1 - Beskrivning av datakällan syfte, innehåll och sammanhang

Denna del syftar till att beskriva källans bakgrund, nuläge och framtid. Beskrivningen förväntas ge en förståelse för datakällans innehåll och egenskaper, vilket bidrar till underlag för beslut om källan ska godkännas samt hur källan bäst ska beredas och tillämpas. Punkterna nedan ska ses som ett smörgåsbord där allt inte alls behöver vara relevant i en specifik utvärdering. Delen ”Innehåll” är förmodligen den som kan vara svårast, speciellt vid sämre dokumenterade källor. Det är inte alls säkert att den delen bäst beskrivs av en mättekniker (snarare än t.ex. ämnesexpert, metodstatistiker eller dataanalytiker), men för att inte ”hamna mellan stolarna” finns det tills vidare finns det med även här.

De metoder som kan användas för att uppfylla punkterna nedan är genomgång av dokumentation och intervjuer av sakkunniga.

Metodinnehåll

Källans syfte för ägaren.

- Vad fyller källan och de aktuella datafälten för funktion för ägaren? Är de centrala eller perifera?
- Hur använder ägaren själv källan?
- Vad blir konsekvenserna av kvalitetsbrister för ägaren?
- Finns det lagar, förordningar och policyer att förhålla sig till för ägaren?

- Vad var behovet och syftet bakom att källan skapades?
- Vad är syftet bakom vägvalen i källans design?
- Vad är ägarens framtida planer med källan (t.ex. finns planerade ändringar med implikationer för SCB:s användning?)

Källans sammanhang:

- Hur ser incitamenten ut för personer som ger input till källan (finns det t.ex. motiv för underrapportering för att undvika beskattning?)
- Finns det omvärldsförändringar relevanta för källan, exempelvis lagändringar som medför att data kan ha olika innehåll för olika tidsperioder eller att vissa variabler blir utdaterade.

Innehåll:

- Beskriv datakällan översiktligt.
- Finns det metadata eller andra beskrivningar av datakällans innehåll vad gäller variabler, objekt eller population?
- Vad innehåller källan för:
    - Objekt? Beskriv och definiera.
    - Variabler? Beskriv och definiera.
    - Population? Beskriv och definiera.
    - Geografiskt område? Beskriv och definiera.
    - Tidpunkter och perioder? Beskriv och definiera.
- Är källan fullständigt eller urval ur en större källa? Om det sistnämnda, hur har urvalet gått till och är det representativt?
- Är källan självständig eller delvis baserat på andra källor? Om det sistnämnda, vad karaktäriserar i dessa källor?
- Inkluderar källan känsliga (person)uppgifter?

Tidigare erfarenheter, användning och arbete med källan:

- Vilken kunskap finns om källans kvalitet? Exempelvis vad gäller täckning, mätning och bortfall?
- Hur bedömer ägaren kvaliteten i källan?
- Vilket kvalitetsarbete har bedrivits och bedrivs?
- Har andra använt källan och vilka erfarenheter finns i sådana fall?

Framtid

- Hur ofta uppdateras källan och sker korrigeringar i efterhand?
- Hur ser framtiden ut vad gäller källan och ägaren (t.ex. finns det inplanerade utvecklingsinsatser, introducering av ny teknik etc. som påverkar källans innehåll)?

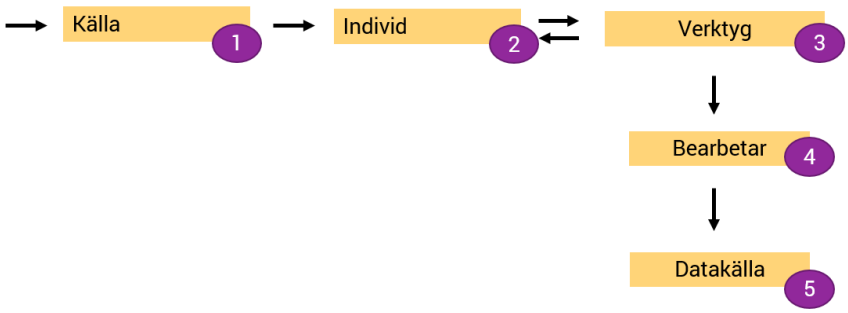
Del 2 - Datagenerering i källa

Denna del handlar om att både översiktligt och detaljerat beskriva och visualisera datagenereringen i källan (här används ”datagenerering” i bred betydelse och inkluderar även den bearbetning (t.ex. kodning) som eventuellt sker internt i källan). Utgångspunkten i arbetet är ett flödesdiagram över de relevanta stegen i processen. För källor som, i sin tur, baseras på många olika typer av källor kan det bli ett komplicerat schema och lämplig nivå måste väljas.

Målet med denna del är att ta fram en processmodell och en övergripande beskrivning av respektive relevant steg. Beroende på vad teamet som arbetar med källan bedömer vara nödvändigt kan det även bli aktuellt med fördjupningar kring ett eller flera specifika steg.

De metoder som kan användas för att uppfylla punkterna nedan är genomgång av dokumentation och intervjuer av sakkunniga. Om källan baseras t.ex. på blanketter eller digitala system för manuella bearbetningar kommer det även att vara aktuellt med traditionellt mättekniskt arbete (t.ex. granskning av blankett och digitalt verktyg). Flera av punkterna nedan, särskilt de som innefattar databearbetning, kommer vid fördjupning att kräva en metodstatistiker. Mätteknikernas roll skulle snarare kunna vara att uppmärksamma förekomst och beskriva tillräckligt utförligt för att bedöma om fördjupning krävs.

EXEMPEL



Figur 1 Exempel på övergripande processmodell

Figur 1 beskriver en övergripande processmodell som innehåller 5 steg och avslutas i den datakälla som SCB ska ta ställning till (alla steg sker alltså utanför SCB, t.ex. hos annan myndighet). Det är ett exempel på en process, inte en principiell modell. Processen i exemplet utgörs av att en individ utgår från en källa (1), interagerar med ett digitalt verktyg för att föra in uppgifter i källan (2,3). Uppgifterna bearbetas (4) för att slutligen resultera i datakällan av intresse (5). Källan i steg 1 kan vara t.ex. personkontakt med myndigheten (t.ex. Försäkringskassan eller Arbetsförmedlingen), självrapporterade uppgifter från person inloggad på myndighetens webbplats eller uppgifter från en annan datakälla

inom eller från annan myndighet. Beroende på vad det är för källa kan det eventuellt även ligga tidigare steg som påverkar utfallet i 1 (t.ex. tidigare kontakter). I mer komplicerade källor kan en avgränsning behöva göras. I steg 2 hanterar en individ uppgiften. De enklaste fallen av hantering är när individen endast överför en redan existerande uppgift från källan (steg 1) till verktyget (steg 3). Mer avancerad hantering är när en individ gör en bedömning av underlaget (t.ex. personkontakt) eller behöver omvandla uppgifterna från underlaget för att passa formatet i verktyget. Bearbetningen i steg 4 skulle kunna vara automatisk och manuell kodning, t.ex. av yrke. Den källa som är av intresse för SCB (5) föregås därmed av en datageneringsprocess som vi behöver ha insikt i för att kunna ha tilltro till och tillämpa källan.

Metodinnehåll

I de punkter där annan kompetens nämns i parentes förväntas samarbete krävas (inte nödvändigtvis med mätteknikern som huvudansvarig).

Input

- Gör ett flödesschema och beskriv översiktligt datagenereringen och den eventuella databearbetning som sker i källan.
- Vad är den datagenererande mekanismen? Är det t.ex. att en individ lämnar uppgifter eller gör en bedömning? Att vi skrapar en webbplats? Att källan registrerar en händelse (t.ex. en transaktion eller gps)?
- Är det en eller flera mekanismer inblandade i data-genereringen? Beskriv dessa.
- Beskriv förutsättningarna för respektive mekanism. Om det t.ex. handlar om att en individ fyllt i en blankett, vilka fält ingår och finns det kontroller? Beskriv eventuella risker för fel.
- Vilka manuella och automatiska moment ingår i datagenereringen?
- Finns det annan dokumentation (rapporter, processdata) som kan bidra tillförståelsen av datagenereringen?

Databearbetning i källa

- Får vi rådata eller har ägaren till källan gjort bearbetningar?
- Förekommer det bearbetningar som korrigeringar, imputering eller kodning? Om ja, vilka bearbetningar görs och i vilken omfattning?
- Lägg till bearbetningsstegen i flödesschemat
- Beskriv villkoren och övergripande om metoderna för bearbetningarna (involverar metodstatistiker och ämne). Bedöm och beskriv riskerna för fel ur ett mättekniskt perspektiv.
    - Förekommer dubbletter? När i datagenereringen uppkommer en dubblett? Varför uppkommer de?

    - Förekommer kodning? I sådana fall, vad kodas och på vilket sätt?
    - Förekommer imputering? I sådana fall, vad imputeras i vilken omfattning?
    - Förekommer härledda variabler? I sådana fall hur?
    - Förekommer andra bearbetningar eller korrigeringar?
- Vad är ägarens egna erfarenheter av bearbetningarna?
- Finns det kända tekniska problem, svagheter eller händelser t.ex. uppehåll under den aktuella tidsperioden?
- Frågorna om dubbletter, kodning och imputering etc. utreds, vid behov, vidare av metodstatistiker. (se t.ex. kvalitetskriterier för digital data)

Del 3 - Tillämpning av källa för statistiskt ändamål

Metodinnehåll

Denna del handlar om att pröva datakällan mot det tänkta statistiska ändamålet.

- Beskriv det aktuella ändamålet vad gäller population och variabler.
- Hur ser överensstämmelsen ut mellan källans population och målpopulationen för ändamålet? (involverar metodstatistiker och ämne)
- Hur ser överensstämmelsen ut mellan definitionerna av relevanta variabler i källan och målvariabler för statistiken? Om överensstämmelsen inte är fullständig, beskriv skillnaderna. (kan involvera metodstatistiker och ämne)
- Beskriv eventuella mättekniska aspekter av möjligheter till länkning till basregister.
- Hur ser överensstämmelsen ut mellan referensperioderna för källan och statistiken. (kan involvera metodstatistiker och ämne)
- Hur förhåller sig källans variabler till eventuella SCB standarder? (kan involvera metodstatistiker och ämne)
- Finns det andra tänkbara ändamål och användningsområden för SCB? (i den senaste versionen av godkännandeprocessen är detta framlyft. Det behöver troligen därför vara en egen utvärdering och inte en punkt i den mättekniska granskningen men står tills vidare med i punktlistan för att inte missas).

Del 4 - Mättekniska bedömningar

Detta avsnitt ska ses som sammanfattande bedömning, att redovisa och dra slutsatser av de viktigaste resultaten i föregående avsnitt. Det är inte självklart hur mätfel ska definieras när det gäller nya datakällor. I huvudsak kan mätfel uppstå i minst två lägen: 1) direkt i källan, att källan misslyckas med att fånga den uppgift den syftar till (t.ex. på

grund av att klient eller handläggare gör fel eller på grund av tekniska problem); 2) i tillämpningen av källan (t.ex. att variablernas definitioner inte fullständigt överensstämmer). Detta kallas även validitetsfel.

Båda typerna förekommer nedan.

Metodinnehåll

Mätfel i källan

- Hur definieras mätfel i källan? (involverar metodstatistiker och ämne)
- Vad är mätteknikerns bedömning av mätfel i källan vad gäller omfattning?
- Vad är mätteknikerns bedömning av mätfel i källan vad gäller orsak?
- Finns behov av att utvärdera mätfel ytterligare? (involverar metodstatistiker och ämne)
- Har mätfelen utvärderats på annat sätt än mätteknikerns bedömning (t.ex. kvantitativ jämförelse med annan källa, etablerad kunskap etc)?
- Hur bör mätfelen hanteras? (involverar metodstatistiker och ämne)

Mätfel - tillämpning

- Vad är mätteknikernas bedömning av mätfel (omfattning eller orsak) om källan skulle användas för det aktuella syftet? (involverar metodstatistiker och ämne).
- Finns det möjligheter att hantera eventuella mätfel? (involverar metodstatistiker och ämne).

Saknade värden

- Hur definieras saknade värden (saknade värden i källan)? (involverar metodstatistiker och ämne)
- Vad är mätteknikerns bedömning av saknade värden i källan vad gäller orsak?
- Har saknade värden och dess orsaker och effekter utvärderats på annat sätt än mätteknikerns bedömning (t.ex. kvantitativ jämförelse med annan källa, etablerad kunskap etc)?
- Hur bör saknade värden hanteras? (involverar metodstatistiker och ämne)

Referenser

Jansson, I. & Japec, L. (2023). Kvalitetskriterier för statistik baserad på digitala data – vägledning.